Unsupervised Learning in Python

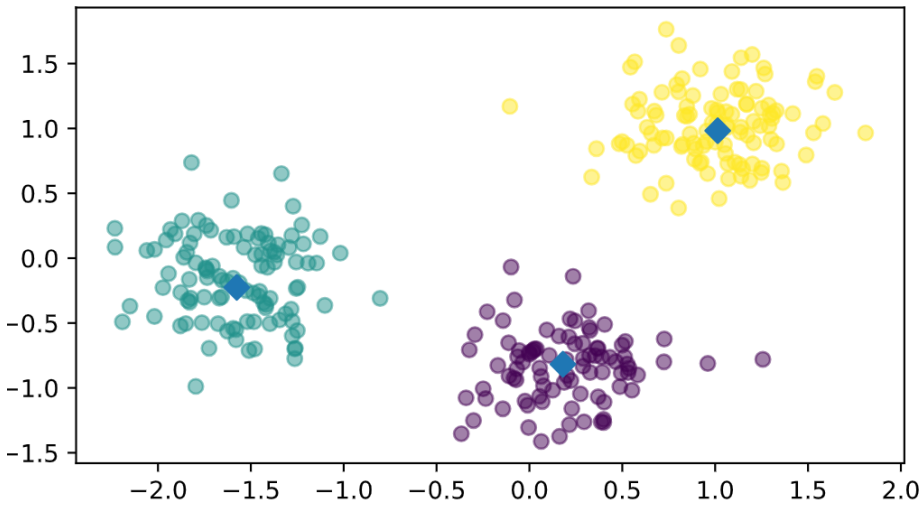
1). Clustering for Dataset Expploration:

a). Clustering 2D Points

```
# Import KMeans
from sklearn.cluster import KMeans

# Create a KMeans instance with 3 clusters: model
model = KMeans(n_clusters=3)

# Fit model to points
model.fit(points)

# Determine the cluster labels of new_points: labels
labels = model.predict(new_points)

# Print cluster labels of new_points
print(labels)
```

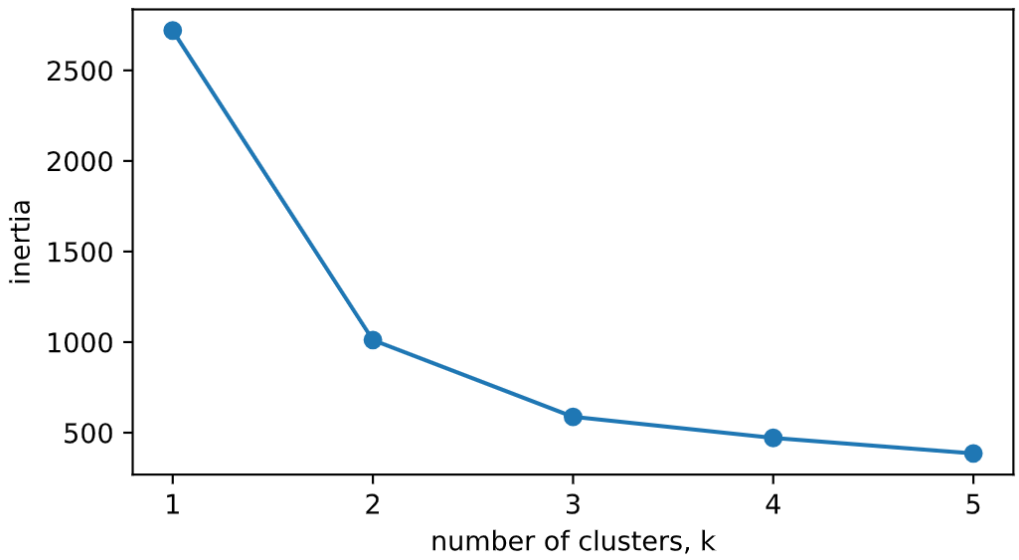
b). Inspect your clustering

```
# Import pyplot
import matplotlib.pyplot as plt

# Assign the columns of new_points: xs and ys
xs = new_points[:,0]
ys = new_points[:,1]

# Make a scatter plot of xs and ys, using labels to define the colors
plt.scatter(xs,ys,c=labels, alpha=0.5)

# Assign the cluster centers: centroids
centroids = model.cluster_centers_

# Assign the columns of centroids: centroids_x, centroids_y
centroids_x = centroids[:,0]
centroids_y = centroids[:,1]

# Make a scatter plot of centroids_x and centroids_y
plt.scatter(centroids_x, centroids_y, marker='D',s=50)
plt.show()
```



c). How many Clusters of grains

```
ks = range(1, 6)
inertias = []

for k in ks:
    # Create a KMeans instance with k clusters: model
    model=KMeans(n_clusters=k)

    # Fit model to samples
    model.fit(samples)

    # Append the inertia to the list of inertias
    inertias.append(model.inertia_)

# Plot ks vs inertias
plt.plot(ks, inertias, '-o')
plt.xlabel('number of clusters, k')
plt.ylabel('inertia')
plt.xticks(ks)
plt.show()
```



d). <u>Evaluating the grain clustering</u>

```
# Create a KMeans model with 3 clusters: model
model = KMeans(n_clusters=3)

# Use fit_predict to fit model and obtain cluster labels: labels
labels = model.fit_predict(samples)

# Create a DataFrame with labels and varieties as columns: df
df = pd.DataFrame({'labels': labels, 'varieties': varieties})

# Create crosstab: ct
ct = pd.crosstab(df['labels'],df['varieties'])

# Display ct
print(ct)
```

<script.py> output:

varieties	Canadian wheat	Kama wheat	Rosa wheat
labels			
0	2	60	10
1	0	1	60
2	68	9	0

e). Scaling Fish dta for clustering

```
# Perform the necessary imports
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Create scaler: scaler
scaler = StandardScaler()

# Create KMeans instance: kmeans
kmeans = KMeans(n_clusters=4)

# Create pipeline: pipeline
pipeline = make_pipeline(scaler,kmeans)
```

f). Clustering Fish Data:

```
# Import pandas
import pandas as pd

# Fit the pipeline to samples
pipeline.fit(samples)

# Calculate the cluster labels: labels
labels = pipeline.predict(samples)

# Create a DataFrame with labels and species as columns: df
df =pd.DataFrame({'labels':labels, 'species':species})

# Create crosstab: ct
ct = pd.crosstab(df['labels'],df['species'])

# Display ct
print(ct)
```

<script.py> output:

species	Bream	Pike	Roach	Smelt
labels				
0	33	0	1	0
1	0	17	0	0
2	0	0	0	13
3	1	0	19	1

g). Clustering stocks using KMeans

```
# Import Normalizer
from sklearn.preprocessing import Normalizer

# Create a normalizer: normalizer
normalizer = Normalizer()

# Create a KMeans model with 10 clusters: kmeans
kmeans = KMeans(n_clusters=10)

# Make a pipeline chaining normalizer and kmeans: pipeline
pipeline = make_pipeline(normalizer,kmeans)

# Fit pipeline to the daily price movements
pipeline.fit(movements)
```

h). <u>Which stocks move together?</u>

```
# Import pandas
import pandas as pd

# Predict the cluster labels: labels
labels = pipeline.predict(movements)

# Create a DataFrame aligning labels and companies: df
df = pd.DataFrame({'labels': labels, 'companies': companies})

# Display df sorted by cluster label
print(df.sort_values('labels'))
```

<script.py> output:

	companies	labels
29	Lookheed Martin	0
36	Northrop Grumman	0
4	Boeing	0
33	Microsoft	1
23	IBM	1
11	Cisco	1
47	Symantec	1
24	Intel	1
51	Texas instruments	1
50	Taiwan Semiconductor Manufacturing	1
56	Wal-Mart	2
25	Johnson & Johnson	2
27	Kimberly-Clark	2
28	Coca Cola	2
38	Pepsi	2
41	Philip Morris	2
40	Procter Gamble	2

9	Colgate-Palmolive	2
39	Pfizer	2
35	Navistar	3
57	Exxon	3
32	3M	3
53	Valero Energy	3
43	SAP	3
44	Schlumberger	3
0	Apple	3
8	Caterpillar	3
10	ConocoPhillips	3
13	DuPont de Nemours	3
12	Chevron	3
26	JPMorgan Chase	4
1	AIG	4
55	Wells Fargo	4
58	Xerox	4
3	American express	4
5	Bank of America	4
18	Goldman Sachs	4
16	General Electrics	4
52	Unilever	5
6	British American Tobacco	5
49	Total	5
46	Sanofi-Aventis	5
42	Royal Dutch Shell	5
19	GlaxoSmithKline	5
37	Novartis	5
15	Ford	6
45	Sony	6
7	Canon	6
48	Toyota	6